Väyrynen, Pertti & Seppänen, Tapio. WordNet:n hyödyntäminen tiedonhaussa. (Using WordNet in Information Retrieval).

Abstract

One of the many problems in Information Retrieval (IR) is that the information seeker does not know exactly what he or she is searching for. When the topic of a query is common, typically expressed in one word, WordNet and its co-ordinate terms can be of help to expand the original query word into related search topics/words.

Kirjoittajat työskentelevät Oulun yliopiston MediaTeam-tutkimusryhmän Kieli- ja audioteknologian kompetenssitiimissä. Lisätietoja osoitteesta http://www.mediateam.oulu.fi?/lang=fi

Osoite

MediaTeam Oulu, Oulun yliopisto
Erkki Koiso-Kanttilan katu 3
90014 Oulun yliopisto
s-posti: pertti.vayrynen@ee.oulu.fi
puh. (08) 553 2789

1. Johdanto

Eräs keskeinen tiedonhaun ongelma on se, että tiedonhakija ei aina tarkkaan ottaen tiedä, *mitä hän itse asiassa hakee* (Smeaton 1997:117). Käyttäjän kannalta helpoin tapa tehdä haku olisi ehkä tehdä haku hyvin yleisellä tasolla, jolloin hakutarve tyypillisesti ilmaistaan yhdellä hakusanalla. Käytännössä tämä tuottaa lähes aina liiaksi osumia.

Jos haun aihe on laaja tai liian yleisesti määritetty kuten esim. *saastuminen* (pollution) hakutopiikkina, niin siihen liittyy yleensä useita eri aspekteja, jotka pitäisi kyetä rajaamaan tarkemmin paremman hakutuloksen saamiseksi.

Ensimmäisenä mieleen tulevia aspekteja saattaisivat olla esim. *saastuttavat aineet* tai *saastumisen seuraukset*, esim. paljas maaperä, joita kuvaavia termejä voisi käyttää uusina hakusanoina, jos nämä termit saisi jostain kätevästi esille. Sinällään paljas maaperä saattaa toki johtua muistakin syistä, joiden todellinen syy saattaa kuitenkin olla saastuminen.

Käyttäjän kannalta ideaaliratkaisu saattaisi olla se, että hän voisi määritellä haun ensin hyvin yleisellä tasolla käyttäen yhtä ainoaa hakusanaa, esim. hakutopiikin *saastuminen* osalta hakusana olisi englannin kielessä *pollution*, ja jokin toinen sovellus esittäisi listan sanoja, jotka kuvaavat muita hakutopiikin eri aspekteja kuten esim. *väestökato* (depopulation) eli saastumisesta aiheutuva väestökato saastuneelta alueelta, mikä saattaisi olla se käyttäjää itse asiassa kiinnostava tiedonhaun aihe tai laajemman haun aspekti, josta hän ei ollut alun perin itse tietoinen tai ei huomannut sen liittyvän alkuperäiseen laajempaan hakuun.

Englannin kielen osalta WordNet (Fellbaum, 1991) ja sen *koordinaattitermit* (co-ordinate terms) näyttäisivät toimivan hyödyllisenä tiedon lähteenä, jonka avulla voi saada esiin joitakin haun aihetta tarkemmin rajaavia termejä silloin kun haku on tehty hyvin yleisellä tasolla. WordNetissä koordinaattitermit viittaavat substantiiveihin ja verbeihin, joilla on sama hypernymi eli yläkäsite. Koska koordinaattitermit esiintyvät tekstissä tyypillisesti joko tekstin alussa tai lopussa, jossa tekstin alussa esitellään yleinen propositio tai tekstin lopussa tehdään yleistys edellä olevan tekstin perusteella, koordinaattitermien käytöllä hakusanoina saattaisi olla myös merkitystä topiikin tunnistuksessa tiedonhaussa. Tämän artikkelin tarkoituksena on esitellä WordNet:ä sekä osoittaa esimerkinomaisesti, millaisia haun rajaus- tai laajennusmahdollisuuksia WordNet:in koordinaattitermit mahdollistavat yleisille hakutopiikeille kuten *saastuminen* tai *demokratia*.

2. WordNet

WordNet kuvaa sanojen välisiä semanttisia suhteita kuten esim. synonymiaa tai antonymiaa. WordNet on laaja manuaalisesti koottu tietokonepohjainen leksikko, josta löytyy lisätietoa lähteestä Beckwith et. al (1991). WordNet:n tarkoitus on kuvata käyttöalueesta riippumatonta leksikaalista tietoa perustuen joihinkin periaatteisiin, joiden mukaan ihmistenkin psykologinen leksikko on todennäköisesti rakentunut.

WordNet:n versio 1.6. sisältää leksikaalista tietoa noin 64000:sta substantiivista, verbistä ja adjektiivista. Toisin kuin monet muut sanastot WordNet on kehitetty synkronisesta näkökulmasta, ei diakronisesta näkökulmasta kuten tavallista (McMahon & Smith 1995, 367). Suhteellisen tavanomaisten semanttisten suhteiden kuten esim. synonymia tai antonymia lisäksi semanttisia suhteita kuten esim. X on eräänlainen Y tai X on osa Y:tä kuvataan myös.

Niin sanotut synsetit (synsets) edustavat käsitteitä WorNet:ssä. Synsetit viittaavat sanajoukkoihin, jotka jakavat saman perusmerkityksen. Tämä mahdollistaa sen, että näitä sanajoukkoja voidaan kuvata yksinkertaisesti synseteillä (Marcus 1994, 499). Seuraavat sanojen väliset semanttiset suhteet ovat saatavailla WordNet:ssä. Taulukko 1 pohjautuu lähteeseen Jurafsky & Martin (2000, 604):

Taulukko 1. Sanojen väliset semanttiset suhteet WordNet:n versiossa 1.6

Semanttinen suhde	**Määritelmä**	**Esimerkki**
Yläkäsite	Käsitteistä yläkäsitteisiin	*breakfast > meal*
Alakäsite	Käsitteistä alatyyppeihin	*meal > lunch*
On jäsen	Ryhmistä niiden jäseniin	*faculty > professor*
Jäsen	Jäsenistä ryhmiin, johon ne kuuluvat	*copilot > crew*
On osa	Kokonaisuudesta osiin	*table > leg*
Osa	Osista kokonaisuuteen	*course > meal*
Antonymia	Vastakohdat	*leader > follower*

3. WordNetin koordinaattitermit

Kuten yllä todettiin, WordNet:in, joka varsinainen tarkoitus on kuvata domeenista riippumattomia sanojen välisiä semanttisia suhteita kuten esim. *synonymia* tai *antonymia*, koordinaattitermit näyttäisivät mahdollistavan hyvin yleisten hakujen, joita tiedonhakija todennäköisesti tekee silloin kun ei oikeastaan tiedä, mitä hän tarkkaan ottaen hakee, tarkemman rajaamisen yleisten hakusanojen kuten esim. *pollution* tietyn merkityksen mukaan. Esimerkiksi hakusanalle *pollution* löytyvät WordNet:ssä versio 1.6 seuraavat koordinaattitermit kolmessa eri merkityksessä:

Merkitys 1: *pollution* 'the state of being contaminated by harmful substances'

-environmental condition (the state of the environment)
-pollution (the state of being contaminated by harmful substances)
-erosion (condition in which the earth's surface is worn away by the action of water and wind)
-deforestation (the sate of being clear of trees)
-depopulation (the condition of having reduced numbers of inhabitants)
-climate, clime (the weather in some location averaged over some long period of time)
-meteorological conditions (the prevailing meteorological conditions as they influence the prediction of weather)

Merkitys 2: *befoulment*, *defilement*, *pollution* 'the state of being polluted'

-dirtiness, uncleanness (the state of being unsanitary)
-dirt, filth, grime, soil, stain, grease (the sate of being covered with unclean things)
-befoulment, defilement, pollution (the state of being polluted)
-griminess, grubbiness (the sate of being grimy)
-sordidness, squalor, squalidness (sordid dirtiness)
-dinginess (discoloration due to dirtiness)
-smuttiness, sootiness (the state of being dirty with soot)

Merkitys 3: *contamination, pollution* ‘the act of contaminating or polluting, including (either intentionally or accidentally) unwanted substances or factors’

-soiling, soilure, dirtying (the act of soiling something)
-staining, spotting, maculation (the act of spotting or staining something)
-contamination, pollution (the act of contaminating or polluting, including (either intentionally or accidentally) unwanted substances or factors)

Kuten yllä olevista koordinaattitermeistä nähdään, esim. merkityksessä 1 koordinaattitermit kuvaavat *saastumisen seurauksia* (erosion, deforestation, depopulation) tai *ilmastonmuutoksia* pitkällä aikavälillä (climate, clime). Merkityksessä 2 koordinaattitermit kuvaavat erilaisia *ympäristön likaisuuteen* liittyviä aspekteja. Merkityksessä 3 taas koordinaattitermit näyttäisivät liittyvän enemmän itse *saastuttamiseen toimintona* liittyviin aspekteihin.

Kuten yllä annetusta esimerkistä nähdään WordNet:n koordinaattitermit tuovat esiin jo aika monipuolisesti saastumisen erilaisia aspekteja, joista käyttäjä voisi valita hakua tarkemmin rajaavia termejä. Toki haun kohteena voi olla myös jokin muu saastumisen aspekti yllä lueteltujen lisäksi, joka voi olla varsinainen haun kohde.

Millaisia koordinaattitermejä saisimme WordNet:stä hakutopiikille *demokratia* sitten?

Merkitys 1. *democracy* (the political orientation of those who favor government by the people of by their elected representatives)

-political orientation, ideology, political theory (an orientation that characterizes the thinking of a group or nation)

-absolutism, totalitarianism, totalism, (the principle of complete and unrestricted power in government)
-….

-hawkishness (any political orientation favoring aggressive polices)

Merkitys 2. *democracy*, republic, *commonwealth* (a political system governed by the people or their representatives)

-political system, form of government (the members of social organization who are in power)
-autocracy, autarchy (a political system governed by a single individual)
-constitutionalism (a constitutional system of government (usually with a written constitution)
- democracy, republic, commonwealth (a political system governed by the people or their representatives)
-hegemony (the domination of one state over its allies)
-oligarchy (a political system governed by few people)
-plutocracy (a political system governed by the wealthy people)
-technocracy (a form of government in which scientists and technical experts are in control)
-theocracy (a political unit governed by a deity (or by officials thought to be divinely guided)

Kuten yllä olevasta esimerkistä nähdään merkityksessä 1 koordinaattitermeillä annetaan lista erilaisista *isemeistä,* kaiken kaikkiaan 29 kappaletta, jotka kuvaavat erilaisia poliittisia aatteita. Näistä käyttäjä voisi valita sen itseään eniten kiinnostavan ismin, jos se sattuisi olemaan kiinnostava tiedonhaun aspekti. Merkityksessä 2 koordinaattitermit kuvaavat erilaisia *poliittisia järjestelmiä,* joita on olemassa tai on joskus ollut olemassa mahdollistaen haun fokusoinnin johonkin kiinnostavaan poliittiseen järjestelmään.

Käsitteellisesti yllä kuvattu hakukyselyjen laajennus WordNet:n koordinaattitermeillä on lähellä hakukyselyjen laajennusta (query expansion) esim. synonyymien avulla, jotka ovat myös saatavana WordNet:stä.

WordNet:n termistöä on kyllä testattu esim. synonyymien osalta hakukyselyjen laajennuksessa vaihtelevin tuloksin. Ei ole tiedossa, onko WordNet:n koordinaattitermejä/hypernymejä testattu hakukyselyjen automaattisessa laajennuksessa, mutta niillä näyttäisi olevan käyttöä myös silloin kun tiedonhakija ei oikein tiedä mitä hän tarkkaan ottaen hakee.

WordNet:n koordinaattitermien (hyponymien ja hypernymien) käytöllä saattaisi olla myös laajempaakin merkitystä tiedonhaussa hakukyselyjen rajausten tai laajennusten ohella, koska ne esiintyvät tekstissä *vain* tietyntyyppisissä virkkeissä: Hypernymit esiintyvät tekstin yleisimmässä virkkeessä. Koska teksti yleensä alkaa yleisellä propositiolla ja päättyy yleistykseen, joka johdetaan edellä olevasta tekstistä, hypernymin sisältävä virke esiintyy joko tekstin alussa tai lopussa (James 1980, 105). Tällöin hypernymien/koordinaattitermien käytöllä hakusanoina saattaisi olla merkitystä myös topiikin paikannuksessa.

4. Päätäntä

Eräs keskeinen tiedonhaun ongelma on se, että tiedonhakija ei tarkkaan ottaen tiedä mitä hän hakee. Ohjelmistojen kuten esim. WordNet, joka on jopa vapaasti imuroitavissa netistä (http://www.cogsci.princeton.edu/~wn/), ja sen koordinaattitermien avulla käyttäjä voi tarkentaa tiedonhaun kohdetta silloin kun hakutopiikki on yleinen kuten esim. *saastuminen* tai *demokratia*. Koska hypernymit/koordinaattitermit esiintyvät vain tietyntyyppisissä virkkeissä joko tekstin alussa tai lopussa, niiden avulla olisi kenties mahdollista paikantaa topiikkeja tiedonhaussa. Uusin WordNet:n versio 1.7.1. B näyttää sisältävän jopa *enemmän* koordinaattitermeja kuin versio 1.6., jonka termistöä käytettiin esimerkkeinä tässä artikkelissa. Koska WordNet:n saa imuroitua oman koneensa näytölle vapaasti, sen avulla voi kätevästi tutkia yleisten hakutopiikkien eri aspekteja.

Viitteet

Beckwith, R., Fellbaum, C., Gross, G., Miller, G. (1991). WordNet: A Lexical Database Organised on Psycholinguistic Principles. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Toimittanut Zernik, U. Lawrence Erlbaum, Hilldale, N.J.

Fellbaum, C. (toim.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

James, C. (1980). *Contrastive Analysis*. Longman.

Jurafsky, D. & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Marcus, M. (1994). New Rrends in Natural Language Processing: Statistical Natural Language Processing. *Voice Communication between Humans and Machines*. Toimittaneet Roe, D. B. & Wilpon, J. G. Washington D.C.: National Academy Press.

McMahon, J. & Smith, F. J. (1995). A Review of Statistical Language Processing Techniques. *Artificial Intelligence Reviews.* Vol. 12. No. 5:347-391.

Smeaton, A. F. (1997). ”Information Retrieval: Still butting heads with natural language processing”. In *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, pp. 115-138. Reprinted as Springer-Verlag Lecture Notes in Artificial Intelligence 1299.